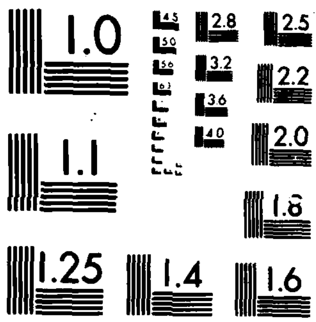
AD-A163 972 ROBUST/RESISTANT TECHNIQUES OF DATA ANALYSIS(U) HARVARD UNIV CAMBRIDGE MASS DEPT OF STATISTICS D C HOAGLIN ET AL. 28 OCT 85 ARO-19236.10-MA 1/1

UNCLASSIFIED DAAG29-82-K-0085 F/G 9/2 NL

END

FILMED

DTIC



MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS 1963 A

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

(2)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
ARO 19236.10-MA	N/A	N/A

4. TITLE (and Subtitle)
Robust/Resistant Techniques of Data Analysis: Final Report

5. TYPE OF REPORT & PERIOD COVERED
final report
1 May 82 - 31 Aug 85

6. PERFORMING ORG. REPORT NUMBER
AR-90

7. AUTHOR(s)
David C. Hoaglin and Frederick Mosteller

8. CONTRACT OR GRANT NUMBER(s)
DAAG 29-82-K-0085

9. PERFORMING ORGANIZATION NAME AND ADDRESS
Department of Statistics, Harvard University
One Oxford Street
Cambridge, MA 02138

10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS

CONTROLLING OFFICE NAME AND ADDRESS
U. S. Army Research Office
Post Office Box 12211
Research Triangle Park, NC 27709

12. REPORT DATE
28 October 1985

13. NUMBER OF PAGES
15

MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)

15. SECURITY CLASS. (of this report)
Unclassified

15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

NA

18. SUPPLEMENTARY NOTES

The view, opinions, and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy, or decision, unless so designated by other documentation.

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

robustness, exploratory data analysis, confirmatory data analysis, critical data analysis, distribution shape, statistical software, analysis of variance

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

Research under this contract focused on three basic problem areas: (1) "critical data analysis" (whose aim is to provide more formal inferences accompanying techniques of exploratory data analysis), (2) distribution shapes that tend to arise in real data, and (3) computer implementation of some advanced techniques in exploratory data analysis. This final report summarizes the research and the major results within each of these three areas.

ARO 19236.10-MA

ROBUST/RESISTANT TECHNIQUES OF DATA ANALYSIS

FINAL REPORT

David C. Hoaglin
and
Frederick Mosteller

28 October 1985

U. S. ARMY RESEARCH OFFICE

Contract DAAG29-82-K-0085
with
Harvard University

APPROVED FOR PUBLIC RELEASE: DISTRIBUTION UNLIMITED.

86 2 7 260

Contents

PROBLEMS AND RESULTS

Research under this contract focused on three basic problem areas: (1) "critical data analysis" (whose aim is to provide more formal inferences accompanying techniques of exploratory data analysis), (2) distribution shapes that tend to arise in real data, and (3) computer implementation of some advanced techniques in exploratory data analysis.

Critical Data Analysis

Topics of interest for our work on critical data analysis included robust/resistant methods in analysis of variance, properties of robust/resistant methods as compared to nonparametric methods, multiplicity and simultaneous confidence, and robust estimation in nonsymmetric situations.

During the period of the contract, our research concentrated primarily on analysis of variance and related models.

In analyzing data that take the form of a two-way layout it is often helpful to consider models that involve both additive and multiplicative terms. One common model of this type decomposes y_{ij}, the value of the response variable in row i and column j, according to

$$y_{ij} = \mu + \alpha_i + \beta_j + \kappa\gamma_i\delta_j + \epsilon_{ij} ,$$

where, when one is fitting by least squares, $\Sigma\alpha_i = \Sigma\beta_j = \Sigma\gamma_i = \Sigma\delta_j = 0$, $\Sigma\gamma_i^2 = \Sigma\delta_j^2 = 1$, and the ϵ_{ij} are uncorrelated. Within this framework we studied some consequences of the nonresistance inherent in least-squares fitting and investigated a robust/resistant approach to fitting such additive-plus-multiplicative models.

When one uses least squares to fit an additive-plus-multiplicative model, a perturbation of the y-value in a single cell can have far greater impact on the fitted value in the same and other cells than is predicted by the formula for leverage in the additive model

$$y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij} .$$

Emerson, Hoaglin, and Kempthorne (1984) derived formulas for generalizations of leverage in particular cases.

To get around such difficulties with least-squares fitting, Emerson and Wong (1985) further developed an approach based on the exploratory technique known as median polish. By making row and column sign changes in the table of additive residuals and then transforming to a logarithmic scale, this approach produces resistant additive-plus-multiplicative fits and can also provide a basis for fitting further multiplicative terms. Hoaglin, Wong, and Emerson (1983) used this procedure as part of a broad framework for resistant diagnosis of interaction in two-way layouts. Also, Emerson, Hoaglin, Tukey, and Wong (1985) illustrated this approach to additive-plus-multiplicative models, together with other resistant techniques, in reanalyzing a classical set of data on the perceived favorableness of 15 adjectives when modified by each of 9 adverbs.

In a related area to analysis of variance, Kempthorne (1984), working in part from a Bayesian viewpoint, devised a procedure for identifying influential groups of observations in multiple regression. By using a direction search to reveal "derivative-influential" data, it offers advantages (especially in computing effort) over other methods.

To provide some inferential support for one frequently used technique of exploratory data analysis, Hoaglin, Iglewicz, and Tukey (1985) carried out an extensive study (both theoretical and empirical) of a class of resistant rules for labeling possible outliers in univariate samples. One main motivation is that, by using measures of location and spread that are themselves relatively insensitive to moderate numbers of sour observations, these rules can avoid most of the problems that many other outlier-detection rules encounter when a sample may contain several outliers. The resistant rules use the lower fourth F_L and upper fourth F_U (approximate quartiles) of the sample to set up cutoffs

$$F_L - k(F_U - F_L) \quad \text{and} \quad F_U + k(F_U - F_L)$$

and label as possible outliers any observations that fall outside these cutoffs. The main rule used in exploratory data analysis has $k = 1.5$ for all sample sizes n. An important aspect of a rule's performance is its "outside rate per sample" (the probability that a sample of n contains at least one "outside" observation). Our work showed that this rule's outside rate per sample ranges roughly from 15 to 35 percent in Gaussian samples of 5 to 50 and generally increases with n. Another characteristic, the outside rate per observation, is roughly 2 to 5 percent around $n = 10$ and decreases as $1/n$ to a Gaussian asymptotic value of 0.7 percent. This finding is of considerable interest, because the outside rate per observation is much higher in small to moderate samples than intuition, based primarily on the population value, had suggested. Hoaglin, Iglewicz, and Tukey also developed (1) a very good theoretical approximation for the outside rate per Gaussian sample that applies to many rules of the above form and (2) a satisfactory approximation, based on the ratio of independent linear combinations of independent exponential variates, for the outside rate per sample in a class of heavier-tailed distributions.

Distribution Shape

One major objective of the research on distribution shape is a better understanding of the variety and characteristics of distributions that arise in actual data, in part as a basis for judging the degree of robustness that various statistical analyses might require in practice. As a framework for studying these questions in continuous data, we have used Tukey's family of g-and-h distributions, which permit more resistant estimates of shape parameters and offer greater flexibility than the traditional third and fourth moments.

In the g-and-h distributions the parameter g controls skewness (departure from symmetry), and the parameter h controls elongation (heavier tails). The basic random variable Y (to which location and scale parameters can be applied) is given, in terms of a standard Gaussian random variable Z, by

$$Y = g^{-1}(e^{gZ} - 1)e^{hZ^2/2} .$$

Estimation of g and h customarily begins with sample quantiles.

Because the behavior of the resistant estimators of g and h had received little attention, Godfrey (1985) studied them in a variety of situations in which the data come from known theoretical distributions, including Gaussian, lognormal, Student's t with small degrees of freedom, and contaminated Gaussian. The sample sizes were 100, 200, 500, and 1000. She found that simple resistant estimators of g and h have distributions very close to Gaussian, appear to be unbiased, and have variances in good agreement with the values predicted by the asymptotic formulas that she derived. The fact that these variances are not especially small confirms the belief that one needs samples of at least several hundred observations to learn much about distribution shape. Her results also suggest that those resistant estimators of g and h are substantially more variable than the corresponding maximum-likelihood estimators.

To illustrate the analysis of distribution shape, as well as to prepare a procedure for later, more routine studies, Godfrey applied the g-and-h techniques to several moderate to large published data sets, ranging in size from 100 to 1500 observations. Godfrey, Hoaglin, and Mosteller began an ongoing program of collecting sizable samples, frequency distributions, and data sets from a variety of sources.

The g-and-h distributions are also valuable in approximating quantiles of non-Gaussian theoretical distributions. Godfrey used this approach to obtain good new approximations for the quantiles of the chi-squared and t distributions.

In other work related to the g-and-h distributions, Hoaglin (1985) described a method for fitting these distributions to binned frequency distributions.

Although discrete distributions pose rather different problems for description of shape, several of the most common families (including Poisson, binomial, and negative binomial) are amenable to flexible resistant checking. Hoaglin and Tukey (1985) substantially improved the Poissonness plot and developed several new techniques for checking the shape of discrete frequency distributions.

Software

Work in this area was designed to make selected advanced techniques of exploratory data analysis more readily accessible for application by implementing them in Fortran. One major product was a set of subroutines that provide almost all the new techniques for diagnosing discrete frequency distributions described by Hoaglin and Tukey (1985).

Also, other aspects of the overall research led to the development of related software, designed for more than casual internal use. The work on outlier labeling, for example, produced an algorithm for evaluating the cumulative distribution function of the ratio of two independent linear combinations of independent exponential random variables. In addition, some software took the form of macros for the Minitab statistical system. These included the singular value decomposition (primarily for fitting additive-plus-multiplicative models by least squares), the biweight location estimator, and biweight polish for two-way tables.

PUBLICATIONS AND TECHNICAL REPORTS

(Reports submitted for publication have been omitted when superseded by a published version.)

David C. Hoaglin, "Discussion" [of the Toolpack Papers by Ford/Hague/Lambert and Osterweil], in 1982 Proceedings of the Statistical Computing Section. Washington, D.C.: American Statistical Association, pp. 30-31.

D. C. Hoaglin, J. D. Emerson, and P. J. Kempthorne, "A Numerical Study of Leverage in Nonlinear Models for Two-Way Tables," 1983 Proceedings of the Statistical Computing Section. Washington, D.C.: American Statistical Association, pp. 143-148.

J. D. Emerson, D. C. Hoaglin, and P. J. Kempthorne, "Leverage in Least Squares Additive-Plus-Multiplicative Fits for Two-Way Tables," Journal of the American Statistical Association, 79 (1984), 329-335.

D. C. Hoaglin, G. Y. Wong, and J. D. Emerson, "Resistant Diagnosis of Interaction in Two-Way Tables." Report AR-22, 27 June 1983.

P. J. Kempthorne, "Identifying Derivative-Influential Groups of Observations in Regression with a Direction Search." Report AR-63, 27 September 1984.

J. D. Emerson, D. C. Hoaglin, J. W. Tukey, and G. Y. Wong, "Some Exploratory Analyses of Adverb-Adjective Data." Report AR-74, 30 May 1985.

D. C. Hoaglin, B. Iglewicz, and J. W. Tukey, "Performance of Some Resistant Rules for Outlier Labeling." Report AR-75, 7 June 1985.

P. Langenberg and B. Iglewicz, "Trimmed Mean $\bar{X}$ and R Charts." Report AR-78, 10 June 1985.

D. C. Hoaglin, F. Mosteller, and J. W. Tukey (Eds.), <u>Exploring Data Tables, Trends, and Shapes</u>. New York: John Wiley & Sons, 1985.

Also 3 chapters in the above book:

J. D. Emerson and G. Y. Wong, "Resistant Nonadditive Fits for Two-Way Tables," pp. 67-124.

D. C. Hoaglin and J. W. Tukey, "Checking the Shape of Discrete Distributions," pp. 345-416.

D. C. Hoaglin, "Summarizing Shape Numerically: The <u>g</u>-and-<u>h</u> Distributions," pp. 461-513.

Participating Scientific Personnel

John D. Emerson (Visiting Associate Professor, Middlebury College)

Katherine Godfrey (Graduate Student)

David C. Hoaglin (Research Associate)

Peter J. Kempthorne (Assistant Professor)

Frederick Mosteller (Professor)

Cleo S. Youtz (Mathematical Assistant)

BIBLIOGRAPHY

Emerson, John D., Hoaglin, David C., and Kempthorne, Peter J. (1984). "Leverage in Least Squares Additive-Plus-Multiplicative Fits for Two-Way Tables," Journal of the American Statistical Association, 79, 329-335.

Emerson, John D., Hoaglin, David C., Tukey, John W., and Wong, George Y. (1985). Some Exploratory Analyses of Adverb-Adjective Data. Memorandum AR-74, Department of Statistics, Harvard University.

Emerson, John D. and Wong, George Y. (1985). "Resistant Nonadditive Fits for Two-Way Tables," in David C. Hoaglin, Frederick Mosteller, and John W. Tukey (Eds.), Exploring Data Tables, Trends, and Shapes. New York: John Wiley & Sons, pp. 67-124.

Godfrey, Katherine (1985). Analysis of Distributional Shape Using g-and-h Distributions. Unpublished Ph.D. thesis, submitted to the Department of Statistics, Harvard University.

Hoaglin, David C. (1985). "Summarizing Shape Numerically: The g-and-h Distributions," in David C. Hoaglin, Frederick Mosteller, and John W. Tukey (Eds.), Exploring Data Tables, Trends, and Shapes. New York: John Wiley & Sons, pp. 461-513.

Hoaglin, David C., Iglewicz, Boris, and Tukey, John W. (1985). Performance of Some Resistant Rules for Outlier Labeling. Memorandum AR-75, Department of Statistics, Harvard University.

Hoaglin, David C. and Tukey, John W. (1985). "Checking the Shape of Discrete Distributions," in David C. Hoaglin, Frederick Mosteller, and John W. Tukey (Eds.), Exploring Data Tables, Trends, and Shapes. New York: John Wiley & Sons, pp. 345-416.

Hoaglin, David C., Wong, George Y., and Emerson, John D. (1983). Resistant Diagnosis of Interaction in Two-Way Tables. Memorandum AR-22, Department of Statistics, Harvard University.

Kempthorne, Peter J. (1984). Identifying Derivative-Influential Groups of Observations in Regression with a Direction Search. Memorandum AR-63, Department of Statistics, Harvard University.

END

FILMED

3-86

DTIC